Density support and intrinsic dimension estimation based on a hierarchical delaunay-type simplicial complex

Catherine Aaron

Density support and intrinsic dimension estimation based on a hierarchical delaunay-type simplicial complex

Catherine Aaron

April 13, 2011

Abstract

Let $X_1, ..., X_N$, $X_i \in \mathbb{R}^D$ be an uniform drawn on a compact $d-$dimensional manifold S with $d \leq D$. Here is suggested a new way to estimate both S and d. The method is based on the computation of a set of simplicial complexes (one for each dimension $d \leq D$) and on an inductive criterion to select the "good" one. Each computed complex is a subcomplex of Delaunay's complex computed using $k-$nearest neighbors restriction and local PCA. A proposition for the k value is given in the first part and the algorithm is detailed in the second part.

1 Introduction

1.1 Intrinsic dimension estimation

Dimension estimation is a challenging problem that has many statistical applications in data analysis, the most obvious applications are dimension reduction methods (for instance isomap [13], LLE [20], HLLE [8] or SOM algorithms [15] all need the initial choice of a dimension), but dimension knowledge is also helpful for modeling problems as in time series [21] or regression [5]. Dimension estimation is also related to other mathematical fields as neural network, signal processing [7] and physics [10] so giving an exhaustive bibliography is vain (for a complete review see Cutler's chapter of [4]). We are only going to deal here with the main ideas of intrinsic dimension estimation.

Mostly we can find two types of dimension estimation methods :

- **Topology based dimension estimation :** They are mainly based on the Hausdorf dimension simplification based on Grasberg and Proccacia's work [19]. Let us define :

$$C(r) = \lim_{N\to\infty} \frac{2}{N(N-1)} \sum_{1\leq i\leq j\leq N} \mathbf{1}_{||X_i - X_j||<r}$$

$$d = \lim_{r\to 0} \frac{lnC(r)}{ln(r)}$$

  Since 1992, the limitation of such a method has been studied in [9] (in this paper the needed N according to the dimension and diameter of the set is studied) but there exists two kinds of limitations with such a method :

  - It is strictly based on uniformly randomized data (and there is little hope to remove this hypothesis)
  - The first $C(r)$ can't be computed (because it is only possible to observe a finite number of points) so if the d formula is applied, the estimated dimension will be 0 (the true dimension of a discrete set). A good value for a small r value must be found or the method has to be improved (as in [18] or [17]) for instance)

- **Local PCA :** A radically (and more statistical) approach is the local PCA's approach [14]. The idea is to compute tangent T space via local PCA and to observe the decreasing of the variance of the projection on $T^{\perp}$. The main problem here is to define well the neighborhood of each point. The present work may be useful for this via the theoretical result in section 2.

1.2 Density support estimation and its topological properties

Density support estimation has a lot of application fields, for instance in medical diagnosis, machine condition monitoring, marketing and econometrics as noticed Biau and Pelletier in [2]. The support density estimation using union of small balls centered on observation has been studied in [1], [2], [12]. Such a method has great asymptotical properties but a huge inconvenient : the topological properties of the estimated support may not be those of the "true" support : for instance the dimension of the estimated support is D the dimension of the embedding $\mathbb{R}^D$ in which the support is and not the dimension of the support. There can be holes in the estimated support that do not exist in the true supports so the estimated homology groups might differ from the "true" one.

The importance of the computation of the homology groups for application (and the methods for it) can be seen in [6] and asymptotical properties of an estimation of Betti numbers can be found in [16]. In [11] it has been applied and a Klein bottle shape has been observed in a real data base.

1.3 The proposed method

This paper's aim is to manage to build a complex on the data to estimate the density support. This complex is expected to have the same dimension and the same topological invariant as the unknown true support of the density. Contrary to [22] the final result is a complex on which the homology can be computed (and not a set of complexes which implies to find "the persistent one").

Section 2 is dedicated to the theoretical search of a k to restrict Delaunay's complex by a $k-$nearest neighbors and not to create an undesirable hole (which could skew the computation of Betti Numbers).

Section 3 presents an algorithm that computes a complex for each supposed dimension d using Delaunay's complex and local PCA.

Section 4 gives an indicator to choose a dimension and so a complex.

Finally section 5 presents some results on simultated data.

2 Majoration of the probability to suppress an inside edge by restricting Delaunay's complex by $k-$nearest neighbor

2.1 Introduction

let $X_1, ...X_N$ be a sample on $\mathbb{R}^D$ in a d dimensional submanifold $S \subset \mathbb{R}^D$ $(D \geq d)$.

One of the goal of this paper is to build a "good" complex that links the points of the sample. S will thus be estimated as the union of all the simplexes.

We choose to use initially Delaunay's complex. Let us denote Delaunay's complex T. It satisfies :

- The $D-$ dimensional simplexes of T are the $(X_{i_1}, ..., X_{i_{D+1}})$ such as the hypersphere circumscribed to the points does not contain any X_j
- The $k-$ dimensional simplexes of T are the $k-$sub-simplexes of the $D-$dimensional simplexes
- The $1-$ dimensional simplexes of T will be called edges of the complex
- **Property :** Delaunay's complex gives an estimation of the convex hull of the sample
- **Corollary :** Since S is not convex it is necessary to remove some simplexes to get a good estimation of S using the complex (see figure 1)

This section focuses on the case $D = d$ and the way to remove simplexes from T to estimate correctly S. We choose the well-known rule of constraining D by $k-$nearest neighbors : the removed simplexes will be those that contain at least one edge $[A, B]$ such as B is not a $k-$nearest neigbor of A and A is not a $k-$nearest neighbor of B.

The choice of a "good" value for k is fundamental to get good results. k must be small enough to respect the local non-convexity of S but large enough to avoid unexpected "holes" (which may be a problem if using results to compute the homology group, for instance).

The following part of this section is dedicated to proving that the probability to remove an inside edge when $D = d$ and the drawn is uniform is majored by :

$$P(N, k) \leq a_d^{k-1}(N-2)^{3/2}\sqrt{\pi/2} + a_d^{N-2}$$

with :

$$a_d = \frac{2^d - 1}{2^d}$$

Unfortunately our definition of an inside edge is a little more restrictive than the creation of an undesirable hole but it gives an approximation (see appendix for a discussion about this definition).

(a) (b) (c)

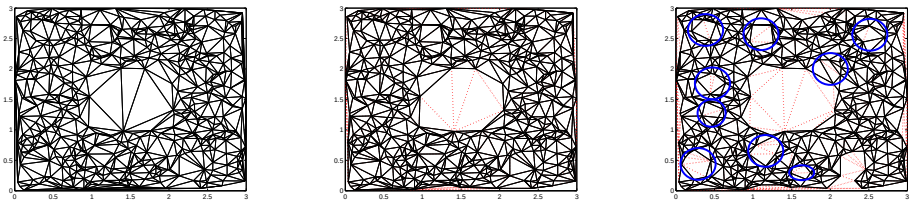


Figure 1: 3 complexes computed on an holed squared sample (500 points $\rightsquigarrow \mathcal{U}([0,3]^2 \setminus [1,2]^2)$). (a) : Delaunay's complexe. (b) : Delaunay's complexe contrained by 42−nearest neighbors. The constrained graph is plain and black ; Delaunay's complex is dashed and red. (c) : Delaunay's complexe contrained by 14−nearest neighbors. The constrained graph is plain and black ; Delaunay's complex is dashed and red. 9 undesirable holes can be observed

2.2 Theoretical Study

In all this part we will assume that $X_1, ..X_N$ is a uniform drawn in $\mathbb{R}^d$ on S a $d-$dimensional bounded manifold. Without loss of generality, we also will assume that $V(S)$ the volume of the manifold is 1.

We will first start by two elementary lemmas :

Lemma 1. *Let us denote :*

- $B_1 = \mathcal{B}(O, r)$
- X *a point of the boundary of* B_1
- O' *such as* $d(O, O') = d(O', X) = r'$
- $B_2 = \mathcal{B}(O', r')$
- $B_1' = B_1 \setminus (B_1 \cap B_2)$

then :

$$\frac{V(B_1')}{V(B_1 \cup B_2)} \leq 1 - \frac{1}{2^d} = a_d$$

Proof. Let first us define $O'' \in [O, O']$ with $d(O, O'') = r/2$ and $B_3 = \mathcal{B}(O'', r/2)$ (see figure 2).

Let us denote $c_d = \pi^{d/2}/\Gamma(1 + d/2)$ (the volume of the unit d-ball).

$B_3 \subset B_1 \cap B_2 \Rightarrow V(B_1') \leq c_d r^d - c_d (r/2)^d$

$B_1 \subset B_1 \cup B_2 \Rightarrow V(B_1 \cup B_2) \geq c_d r^d$

□

Lemma 2.

$$S_n = \frac{1}{n} \sum_{k}^{n-1} \frac{1}{\sqrt{(1/n)(1 - k/n)}} < \pi$$

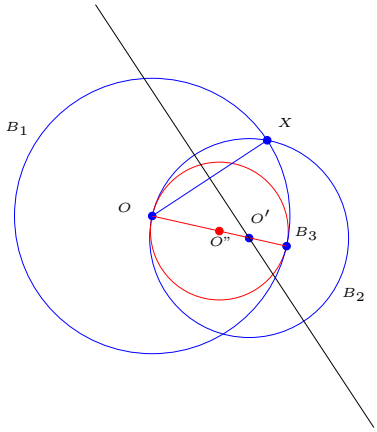


Figure 2: Illustration for lemma 1. Remark for $d > 2$ this graphic corresponds to the projection in the plane containing O, O' and X

Proof. let us denote : $f(x) = 1/\sqrt{x(1-x)}$ defined on $]0,1[$.
We define g_n stepwise by :
$g_n(x) = min\{f(t), t \in [(k-1)/n, k/n]\}$ when $x \in [(k-1)/n, k/n]$
Obviously : $g_n(x) \leq f(x)$ so $\int_0^1 g_n \geq \int_0^1 f = \pi$.
$S_n + f(0.5)/n = \int_0^1 g_n$ (see figure 3) finishes to prove it.

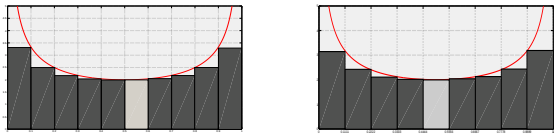


Figure 3: Illustration for lemnm 2 with n even and n odd. The red line represents the function, the step function g_n and the grey area's surface S_n

□

Hypothesis, definitions and notations

- $X = \{X_1, .., X_N\}$ is a uniform sample on a $d-$dimensional manifold S with $V(S) = 1$. For all that follows Delaunay's complex is Delaunay's complex computed on X.
- X_j is the $\hat{k}(X_i, X_j)^{the}$ neighbor of X_i.
- $k^*(X_i, X_j) = min(\hat{k}(X_i, X_j), \hat{k}(X_j, X_i))$.
- If $t = [X_i, X_j]$ is included in Delaunay's complex then $k^*(t) = k^*(X_i, X_j)$.
- an edge $t = [X_i, X_j]$ of Delaunay's complexe is an inside edge if $\mathcal{B}(X_i, d(X_i, X_j)) \cup \mathcal{B}(X_j, d(X_i, X_j)) \subset S$ (see appendix for a discussion of this notion).

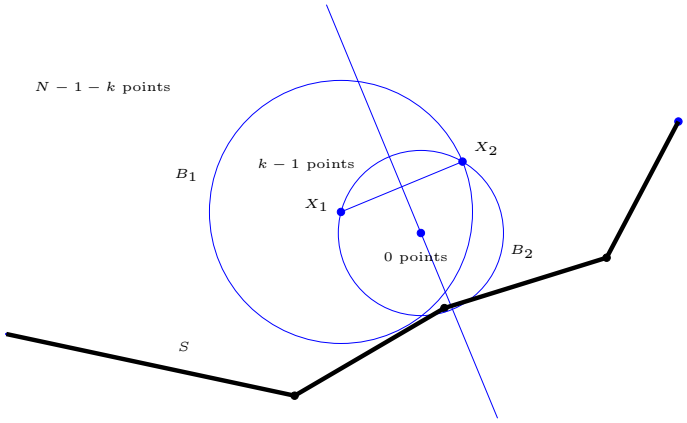


Figure 4: Illustration for lemna 3

Lemma 3. *Considering all the previous hypotheses, definitions and notations, let t be an edge of Delaunay's complexe inside S and $2 \leq k \leq N-2$ then :*

$$P(k^*(t) = k) \leq a_d^{k-1} \frac{\sqrt{N-2}}{\sqrt{2\pi}\sqrt{(k-1)(N-k-1)}} exp\left(\frac{1}{12(N-2)}\right).$$

Proof. Let us first assume that :

- $t = [X_1, X_2]$,
- $\hat{k}(X_1, X_2) = k$.

As t is in Delaunay's complexe there exists an empty ball B_2 such as X_1 and X_2 are on the boundary of B_2. As $\hat{k}(X_1, X_2) = k$ then $B_1 = \mathcal{B}(X_1, d(X_1, X_2))$ contains $k-1$ other points (see figure 4).

Knowing B_2, the probability to get such a configuration is :

$$P^{B_2}(k^*(t) = k) = C_{N-2}^{k-1}(V(B_1 \setminus (B_1 \cap B_2)))^{k-1}(1 - V((B_1 \cup B_2) \cap S))^{N-1-k}.$$

Let us denote $x = V(B_1 \cup B_2)$ and apply lemma 2 :

$$P^{B_2}(k^*(t) = k) \leq C_{N-2}^{k-1} a_d^{k-1} x^{k-1}(1-x)^{N-1-k}.$$

A maximisation of the expresion leads to :

$$P(k^*(t) = k) \leq a_d^{k-1} C_{N-2}^{k-1} \frac{(k-1)^{k-1}(N-k-1)^{N-k-1}}{(N-2)^{N-2}}.$$

Finally, we use the Stirling unequality (for $k \geq 2$ and $k \leq N-2$) to get :

$$P(k^*(t) = k) \leq a_d^{k-1} \frac{\sqrt{N-2}}{\sqrt{2\pi}\sqrt{(k-1)(N-k-1)}} exp\left(\frac{1}{12(N-2)}\right).$$

□

Theorem 1. *Let $X = (X_1, ..., X_N)$ be a uniform sample on a $d-$dimensional manifold S with $V(S) = 1$, the probabilty $P(N, k)$ that an inside edge of Delaunay's complex is suppressed when restricting by the $k-$nearest neighbors graph satisfies : $P(N,k) \leq a_d^{k-1}(N-2)^{3/2}\sqrt{\pi/2}exp\left(\frac{1}{12(N-2)}\right) + a_d^{N-2}$.*

Proof. An inside edge t is suppressed when restricting by the $k-$nearest neighbors graph if $k^*(t) \geq k$ so :

$$P(N,k) = \sum_{j=k}^{N-1} P(k^*(t) = j)$$

Lemma 3 gives

$$P(N,k) \leq \sum_{j=k}^{N-2} a_d^{j-1} \frac{\sqrt{N-2}}{\sqrt{2\pi}\sqrt{(j-1)(N-j-1)}} exp\left(\frac{1}{12(N-2)}\right) + a_d^{N-2}.$$

$$P(N,k) \leq a_d^{k-1} \sum_{j=k}^{N-2} \frac{\sqrt{N-2}}{\sqrt{2\pi}\sqrt{(j-1)(N-j-1)}} exp\left(\frac{1}{12(N-2)}\right) + a_d^{N-2}$$

,

and lemma 2 leads to the conclusion.

□

Corollary 1. *The restriction of the delaunay complex by the $k-$nearest neighbor graph with :*

$$k \geq 1 + \frac{\ln(\varepsilon) - \frac{3}{2}\ln(N-2) - \frac{1}{2}ln(\pi/2)}{\ln(a_d)} = k^0(N, d, \varepsilon)$$

creates an inside edge with a probability :

$$P \leq \varepsilon e^{\frac{1}{12N}} + a_d^{N-2} \sim \varepsilon$$

2.3 Numerical results

2.4 Graphical results

We prensent here 4 graphes, illustrating Delaunay's restriction to $k^0(N, 2, \varepsilon = 0.1)$ for the same holed squares as in figure 1 (N varying in $\{100, 200, 500, 1000\}$). First, it can be observed that there is no undesirable hole creation in the obtained complexes. In the first case $N = 100$ doesn't allow to obtain a complex that respects the topology of the initial set S : there is no hole at all. We think that this is due to the fact that the size of the hole is too small in regard to the density around the hole. The convergence of the computed complex to the density support seems here to occur (the convergence of the algorithm will be studied in a further paper).

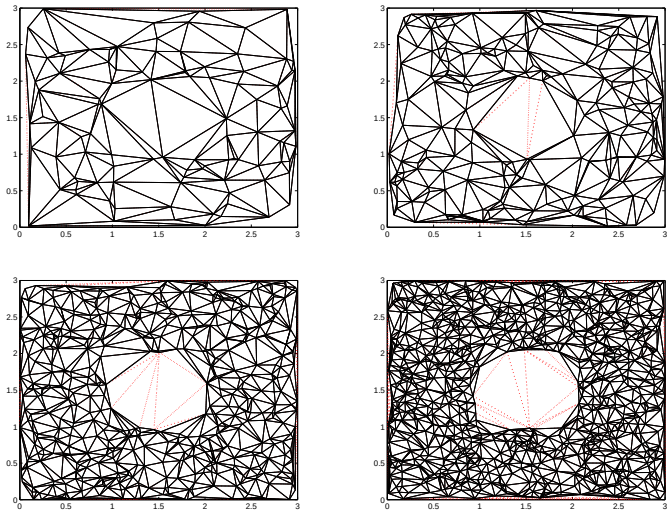


Figure 5: $\varepsilon = 0.1$, $N = 100$, $N = 200$, $N = 500$, $N = 1000$: black edges are those of the restricted Delaunay's complex, dashed red ones are those of the Delaunay's complex. It can be observed that the removed simplex are on the boundary and that the topological properties of the estimated support are the same as the true support since $N \geq 200$

2.5 Simulation on Spheres

Let $K_{d,N}(X_1, ...X_N)$ be the maximum number of neighbors observed on inside edges of Delaunay's complex when the density support is $d-$dimensional and the sample is uniformly distributed on S. Our majoration implies that : $k^0(N, d, \varepsilon) \geq Q_{1-\varepsilon}(K_{d,N})$ (with $Q_{1-\varepsilon}(K_{d,N})$ the $(1-\varepsilon)$ percentile of $K_{d,N}$). It is expected that this majoration is not too big. To compute $K_{d,N}(X_1, ...X_N)$ without boundary effect and so avoiding the problem of looking for the inside edges, the X_i has been uniformly randomized on $d-$dimensional ball $\mathcal{S}_d$ which lies in $\mathbb{R}^{d+1}$. The $d-$dimensional Delaunay's complex is here the boundary of the $(d+1)-$complex. Results of our computation are presented in figure 6 for dimension 1 (1000 draws), 2 (1000 draws) and 3 (500 drawns) and each time $N \in \{100, 200, 300, 500, 1000, 2000\}$. Plain lines represent the "theoretical" k^0 for $\varepsilon \in \{0.1, 0.05, 0.01, 0.005\}$ and dashed lines the simultated percentiles for same ε. It can be observed that :

- The majoration is verified and is not too big
- The growth speed seems to be the right one (for the dimension 3 the $Q_{0.99}$ and $Q_{0.995}$ simulated values might not be good because they are only computed on 500 samples

2.5.1 Some values for k^0 and $\varepsilon = 10^{-2}$

Some numerical values for k^0 are presented in figure 7 : the empty cells correspond to $k^0(N) < N$. Cells in Italics correspond to $k^0 < 2N$ and bold $k^0 > 10N$

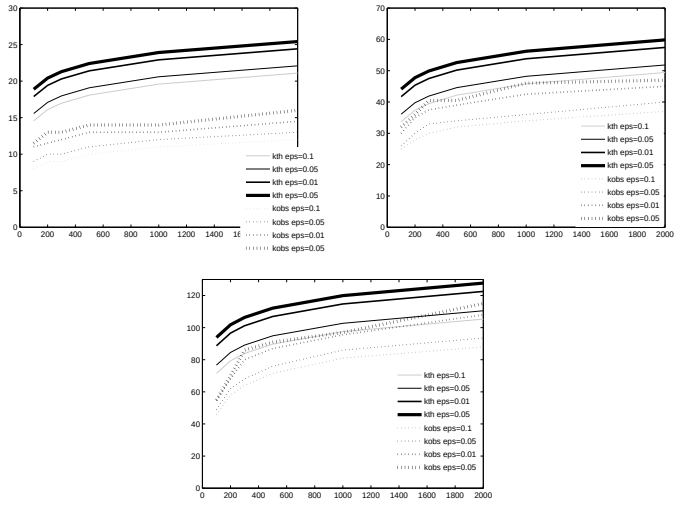


Figure 6: Theoretical (plain) and simulated (dashed) values for k and different values for ε for 1, 2 and 3.

N	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$
50	16	*37*				
100	17	41	*88*			
500	**21**	**50**	106	220	*446*	
1000	**22**	**53**	114	236	479	
2000	**24**	**57**	**122**	252	512	1031
5000	**26**	**62**	**132**	**273**	555	1118
10000	**27**	**65**	**140**	**289**	**588**	1185
20000	**29**	**69**	**148**	**306**	**621**	1251

Figure 7: Some Values for $k^0(N, d, 10^{-2})$.

3 Computation of a simplicial complexe on the data

Let $X = (X_1, ..., X_N)$, $X_i \in R^D$ be a uniform draw on a compact d dimension manifold S with $d \leq D$.

Let us denote T the $D-$dimensional Delaunay's complex of X. We will denote :

$$T = \bigcup_{p \in \{1,..,D\}} \left(\bigcup_i t_i^p \right)$$

with :

- the t_i^D are the $D-$dimensional Delaunay-simplexes : i.e. such as the $D-$ball circumscribed of t_i is empty (doesn't contain any points of X),
- t_i^k is the face of a t_j^{k+1}.

The goal is now to extract of T, T^* a "good" subset of simplexes. By "good" we expect that :

- T^* is a $d-$dimensional simplex with d the true dimension of S,
- T^* gives a correct estimation for S.

For that we will compute D subcomplexes of T (one for each supposed dimension from 1 to D) and the choice of the final complex will be done afterwards.

3.1 Computation of the $D-$dimensional complex

According to section 2 the $D-$ dimensional simplex is the restriction of T by the $k^0(N, D, \varepsilon)-$nearest neighbor graph.

3.2 Computation of the $p-$dimensional complexes for $p < D$

The idea that leads us to the following algorithm is very simple : let us assume that S is a $p-$dimensional smooth manifold. Then locally the manifold and its tangent hyperplan are close and we are going to consider some simplexes of Delaunay's compexe of the local projection on tangent hyperplan. More precisally :

for each point i :

- 1- search $V_i = \{X_{j_1(i)}, ..., X_{j_{k(i)}(i)}\}$ a neighborhood of X_i
  - **Practically :** according to section 2, we propose to use the $k^0(N, d, \varepsilon)-$nearest neighbors of X_i as V_i.
- 2- $H_i^{(p)}$ the hyperplan tangent to S at the point X_i according to the hypothesis that it is $p-$dimensional:
  - **Practically :** using local PCA (PCA on V_i).
- 3- W_i is the set of all the projections of the V_i points on $H_i^{(p)}$.
- 4- compute Delaunay's complex of W_i : $\mathcal{T}_i$.
- 5- keep the $p-$dimensional simplexes of $\mathcal{T}_i$ that satisfy the following properties :
  - i is in the simplex,
  - the simplex is in the set of Delaunay's simplexes T.

The final $p-$dimensional complex will be the union of all the $p-$dimensional simplexes kept in the algorithm.

4 Choice of the intrinsic dimension and complex

It is now needed to choose a complex in the set of all computed complexes (one for each supposed dimension). We suggest here two ways to choose. The first one is only classical local PCA method and the second one is based on the geodesic distances. The local PCA method is not linked to the simplicial approach and can be used beforehand to reduce the set of tested dimension (which can be usueful to reduce the computational time). The second method depends on the simplicial approaches and can be used afterwards to confirm the choice.

4.1 Local PCA

The local PCA method for estimation of the dimension is well known [14]. We here only apply it using the neighborhood according to section 2. Each tested dimension d is associated to a $k^0(N, d, \varepsilon)-$neighborhood and for each point of the sample the eigenvalues of a PCA on its neighborhood can be computed. Boxplotting these eigenvalues can help to choose possible dimensions

4.2 Geodesic distance recognition

Let us suppose that the true dimension is d, then the $k^0(N, d, \varepsilon)-$nearest neighbor graph and the $d-$dimensional complex's graph obtained with our method might both be used to compute approximation of the geodesic distance correctly and so geodesic distances computed using the two graphs might be close. For each dimension, the plot of the geodesic distances computed with nearest-neighbor method and simplicial method will be plotted and a dimension has to be chosen between dimensions that lead to a plot near the bisector.

5 Some examples

5.1 Examples for dimension $D = 2$

We computed here two examples for $D = 2$ ($N = 500$ and $\varepsilon = 0.01$) : the holed square and the circle. For the holed square, let us first look at the local PCA results : they both indicate a dimension 2. Looking at the geodesic distances scatterplots also indicates a dimension 2. The complex associated to the dimension 2 respects the topology of the (known) density support.

For the circle example, the local PCA method and the simplicial method also agree to decide for a dimension 1 (if the hole observed for the dimension 2 had not existed, only the PCA method would have been discriminant to conclude).

5.2 Examples for dimension $D = 3$

We computed here three examples for $D = 3$ ($N = 500$ and $\varepsilon = 0.01$) : the cylinder, the sphere and the spiral.

The sample on the cylinder has been realized as follows : $\theta \rightsquigarrow \mathcal{U}[0, 2\pi]$ $r \rightsquigarrow \mathcal{U}[0.7, 1]$ and $X_3 \rightsquigarrow \mathcal{U}[0, 1]$. $X_1 = cos(\theta)$, $X_2 = sin(\theta)$ and X_3. Local PCA hesitates between dimension 2 and 3 (the third eigenvalues is realy small compared to the two first ones, we can wonder if it is due to a "thin" $3-$dimensional

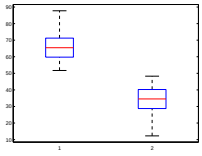
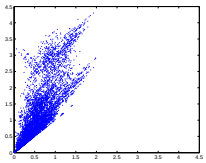
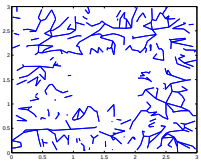
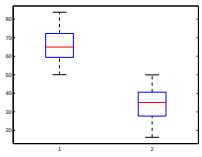
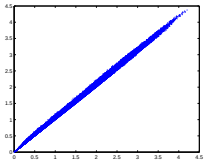
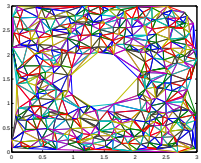
dim	Local PCA	geodesic	complex
1			
2			

Figure 8: example of the holed square

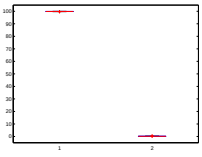
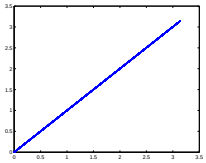
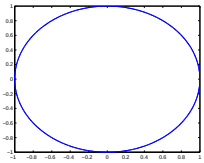
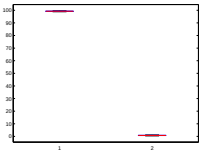
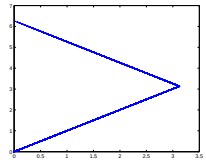
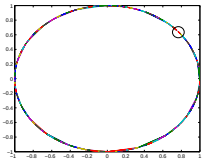
dim	Local PCA	geodesic	complex
1			
2			

Figure 9: example of the circle

S or a non linear $2-$ dimensional S?). Geodesic distance criterion helps to conclude to a dimension $d = 3$.

For the sphere example, the geodesic criterion hesitates between dimension 2 and 3 ; the local PCA helps to conclude to dimension 2.

Conclusions : For all the presented examples (dimension $D = 2$ and $D = 3$) the topological properties of the chosen complexes are the good ones. The choice of the elected complex has to be done according to both local PCA and geodesic approach.

6 Conclusions and persepectives

Our theoretical value for k is coherent with the known good properties that are expected for the choice of k, the $k-$nearest neighborhood (generally expected to satisfy $k \to \infty$ and $k/N \to 0$ ([3]). Fortunately, our majoration in section 2

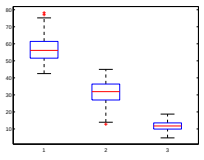
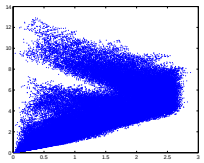
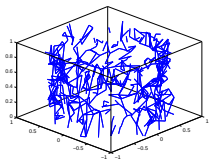
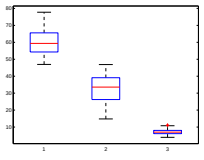
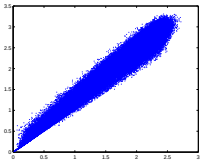
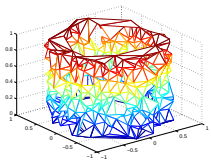
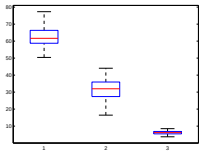
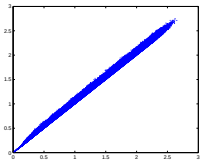
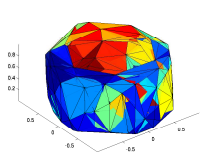
dim	Local PCA	geodesic	complex
1			
2			
3			

Figure 10: example of the cylindre

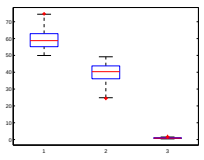
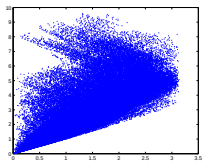
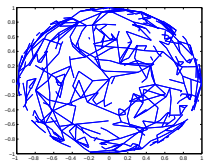
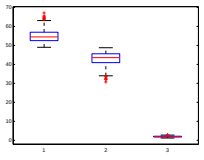
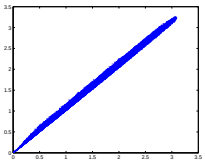
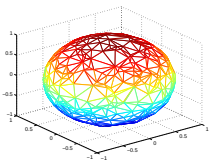
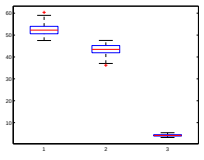
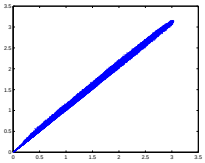
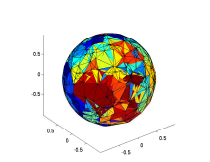
dim	Local PCA	geodesic	complex
1			
2			
3			

Figure 11: example of the ball

Finally for the spiral example it is only the local PCA approach that helps to conclude.

that can be considered as a quite strong majoration is not so far from the simulated value for k. Applications using this value gives quite good results for not

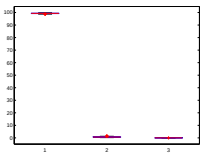
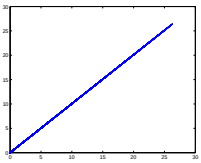
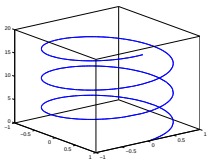
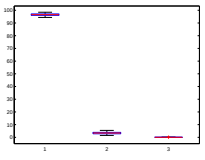
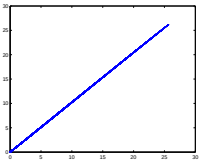
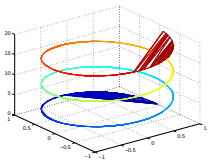
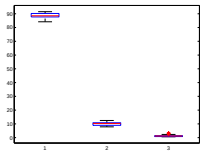
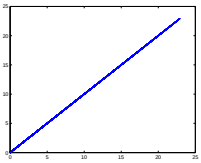
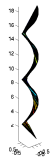
dim	Local PCA	geodesic	complex
1			
2			
3			

Figure 12: example of the spiral

too exotic S but there is still a lot of open questions and possible improvements.

Two main furter theoritical axes can be envisaged :

- On k value to avoid undesirable holes : can we improve the definition of an inside edge to be closer to the hole creation ? Can we prove that $1.5ln(N)/ln(a_d)$ speed is really the good one ? Can our result be extended to non-uniform samples ?
- On the S estimation : does our $\hat{S}_N$ converge ? do Betty number's estimation converges ? what is the speed ? Graphical results of section 5 are encouraging. The intuition is that it is required that $N > c(S)k^0$ with $c(S)$ a constant reflecting the complexity of S.

For the applied part : can the complex building be improved ? can it be adapted to sets S where the local intrinsic dimension is not constant [7]?

References

[1] Alberto Rodrguez-Casal Antonio Cuevas, Ricardo Fraiman. A nonparametric approach to the estimation of lengths and surface areas a nonparametric approach to the estimation of lengths and surface areas. *The Annals of Statistics*, 35:1031–1051, 2007.

[2] Cadre B. Biau, G. and B Pelletier. Exact rates in density support estimation. *Journal of Multivariate Analysis*, 99:2185–2207, 2008.

[3] Crou F. Biau, G. and A. Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research*, 11:687–712, 2010.

[4] P. Lewis-B. Ray P. Brockwell P. Tuan C. Cutler, K. Chan. *DIMENSION ESTIMATION AND MODELS*. World Scientific eBooks, 1993.

[5] B. Cadre and Q. Dong. Dimension reduction for regression estimation with nearest neighbor method. *Electronic Journal of Statistics*, 4:436–460, 2010.

[6] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46:255–308, 2009.

[7] R.; Hero A.O. Carter, K.M.; Raich. On local intrinsic dimension estimation and its applications. *IEEE Transactions on Signal Processing*, 58:650–663, 2010.

[8] D.L. Donoho and C. Grimeslle. Hessian eigenmaps : new locally linear embedding techniques for high-dimensional data. Technical report, standford, 2003.

[9] J.P. Eckmann and D. Ruelle. Fundamental limitations for estimating dimensions and lyapounov exponents in dynamical systems. *Physica*, D56:185–187, 1992.

[10] Yael Almoga Eran Toledoa, Sivan Toledob and Solange Akselroda. A vectorized algorithm for correlation dimension estimation. *Physics Letters A*, 229:375–378, 1997.

[11] V. de Silva G. Carlsson, T. Ishkhanov and A. Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76:1–12, 2008.

[12] David M. MASON Gerard BIAU, Benoit CADRE and Bruno PELLETIER. Asymptotic normality in density support estimation. *Electronic Journal of Probability*, pages 2617–2635, 2009.

[13] V. de Silva J. B. Tenenbaum and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[14] David R.Olsen K. Fukunaga. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 20:176–183, 1971.

[15] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, 1995.

[16] E. Meckes M. Kahle. Limit theorems for betti numbers of random simplicial complexes. Technical report, comptop standford university, 2010.

[17] BoZhangc Mingyu Fana, HongQiaob. Intrinsic dimensionestimationofmanifoldsbyincisingballs. *Pattern Recognition*, 42:780–787, 2009.

[18] Intrinsic Dimension Estimation Using Packing Numbers. Balzs kgl. Technical report, departement d'informatique et de recherche operationnelle universite de Montreal, 2003.

[19] I. Procaccia P. Grassberg. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9:189–208, 1983.

[20] S.T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[21] M. Verleysen, E. de Bodt, and A. Lendasse. *Engineering Applications of Bio-Inspired Artificial Neural Networks*, chapter Forecasting financial time series through intrinsic dimension estimation and non-linear data projection, pages 596–605. Springer Berlin / Heidelberg, 1999.

[22] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33:247–274, 2005.

A Discussion of the inside edge notion

The aim of this appendix is to show that not removing an inside edge (that allows to get the theorem 1) is not so far from not creating an undesirable hole (that is the really interesting notion). The discussion here is not (at all) a proof but helps to understand why, at least for dimension 2, the two notions are close.

Let us remember the definition :

- an edge $t = [X_i, X_j]$ of Delaunay's complexe is denoted "inside" S if $\mathcal{B}(X_i, d(X_i, X_j)) \cup \mathcal{B}(X_j, d(X_i, X_j)) \subset S$

And Let us add two other definitions :

- an edge $t = [X_i, X_j]$ of Delaunay's complexe is denoted "semi-inside" S if $\mathcal{B}(X_i, d(X_i, X_j)) \subset S$ or $\mathcal{B}(X_j, d(X_i, X_j)) \subset S$
- an edge $t = [X_i, X_j]$ of Delaunay's complexe is denoted "not-inside" S if $\mathcal{B}(X_i, d(X_i, X_j)) \not\subseteq S$ and $\mathcal{B}(X_j, d(X_i, X_j)) \not\subseteq S$

Illustrations for these 3 possible cases can be seen in figure 13. Our proof in section 2 can easily be adapted for semi-inside edges because the volume majoration in lemma 1 can be adapted (see figure 13) and because there is at least k point in each presented ball (but the writing is a little more difficult).

Let us now focus on the not-inside edge case and the dimension 2. Let us assume that $t = [X_1, X_2]$ is a not-inside, Delaunay's edge, that satisfies $k^*(t) \geq k$. and that removing t create an undesirable hole. As there is a creation of an undesirable hole t is not on the boundary of the complex. So there are two Delaunay triangles $[X_1, X_2, X_3]$, and $[X_1, X_2, X_3']$. Let us assume that $[X_1, X_2, X_3]$ is the one "closest to the boundary" (see figure 14). $[X_1, X_2, X_3]$ is removed by our algorithm so it is not on the boundary of the complex (otherwise we will not create a hole). Iterating such reasoning leads to the fact that there exists X_4 and X_5 as on figure 14 which satisfies $[X_4, X_5]$ is an edge of Delaunay's complex which is not removed by restrinction to $k-$nearest neighbors. As the construction implies that $d(X_4, X_5) > d(X_1, X_2)$ the fact that $k^*([X_4, X_5]) < k^*([X_1, X_2])$ may be small.

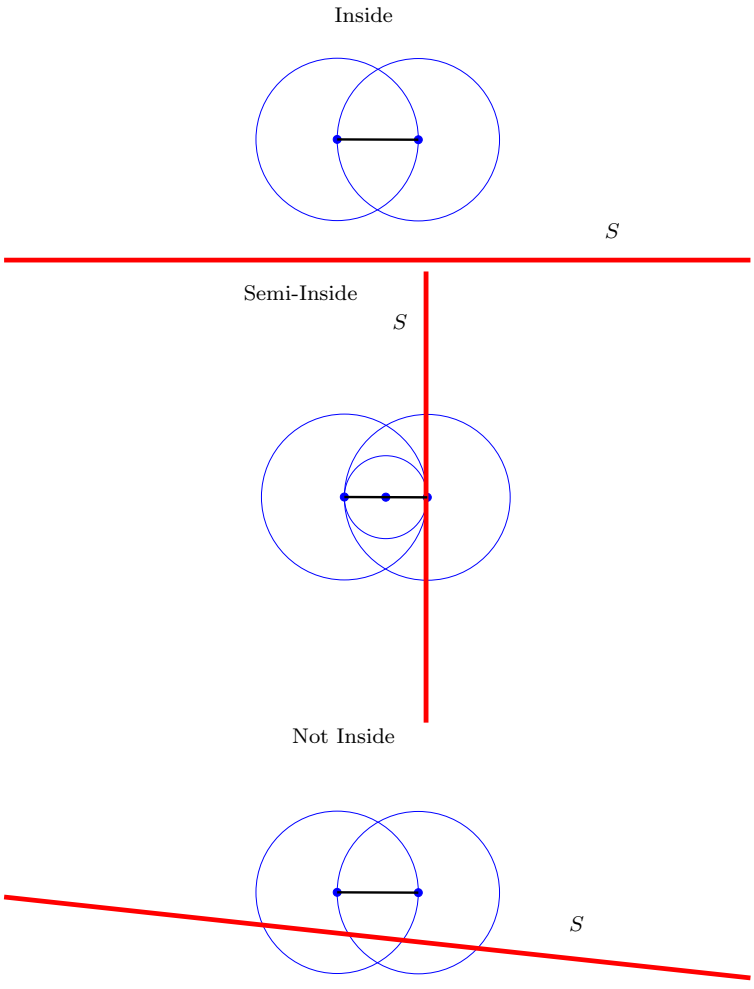


Figure 13: Illustration for inside, semi-inside and not-inside edges

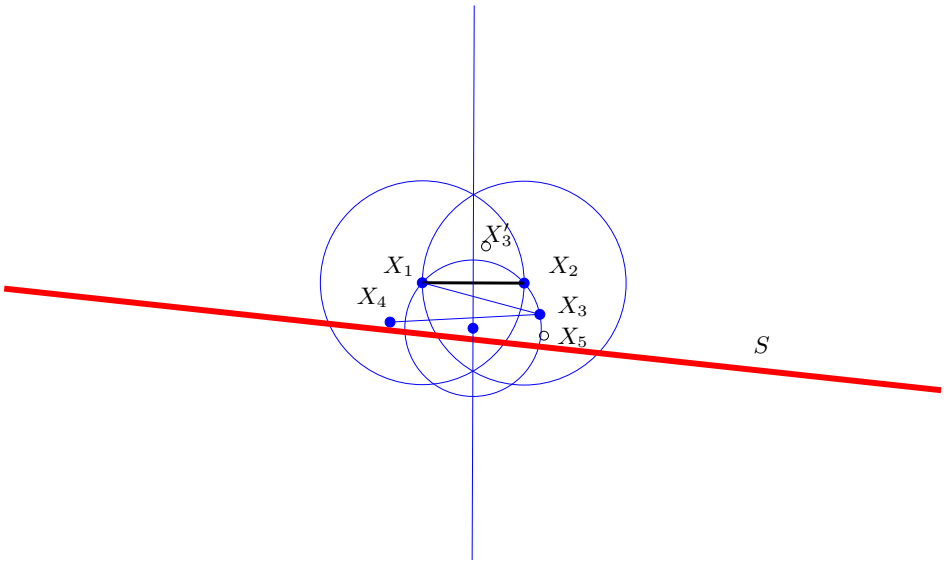


Figure 14: Creating an hole by removing a not-inside edge